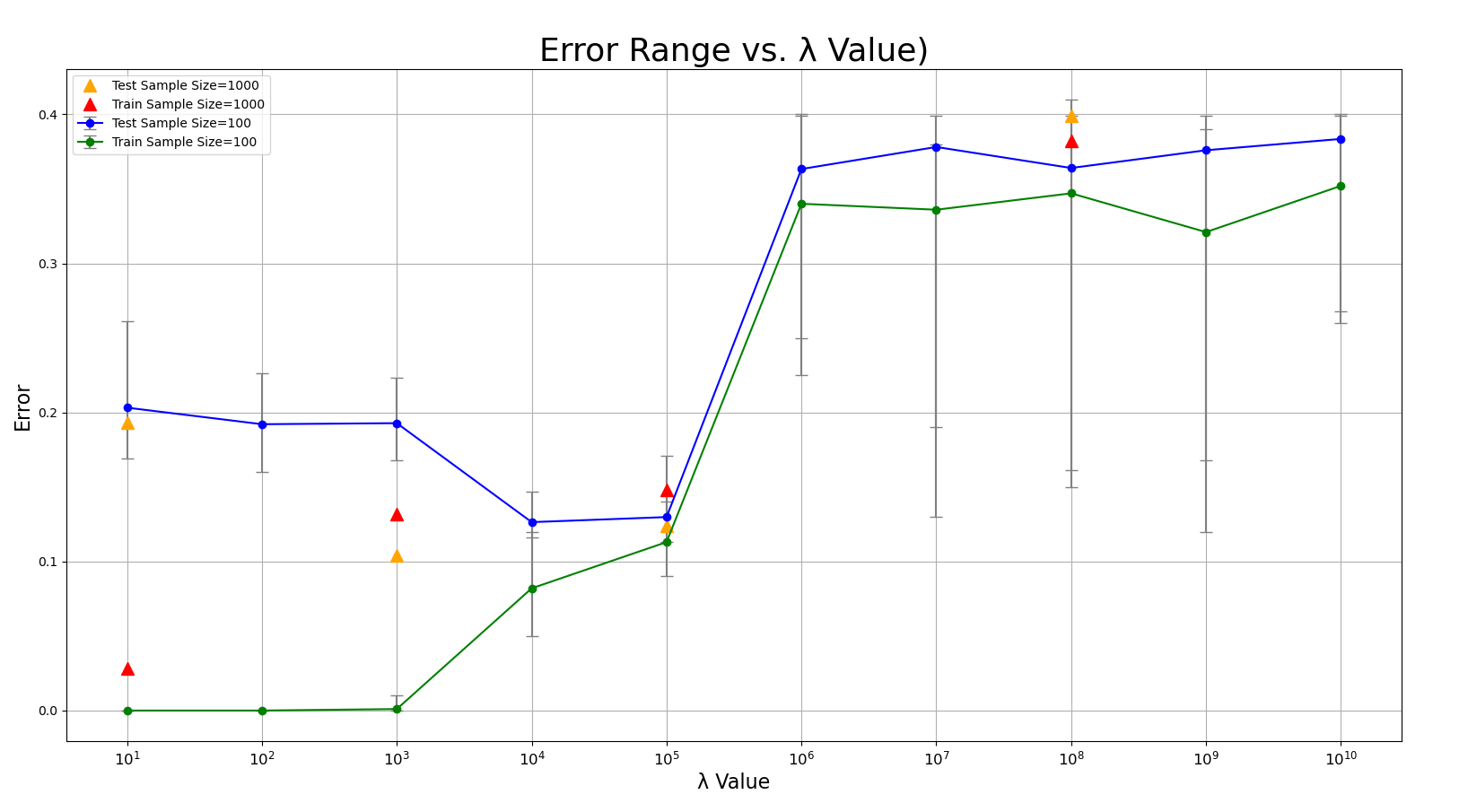
Intro To ML – Linear Predictors

Amit Zulan – 207299033, Dvir Ben Zikri - 315409508

1. Code is added separately.
2. (a) + (b)  
     
   (c)
   * + A smaller sample size should get a lower training error, less examples are easier to separate since they are relatively more scattered (on average), and smaller sample size increases the chances of the samples being linear independent, meaning they can be shattered (i.e. separable regardless of labels). Test error should be lower for higher sample size. Higher sample size avoids over-fitting and generalizes the separator. We can see both are reflected in the plot above for lower .
     + The trend of the training error should be increasing with since lower results in larger which causes a smaller margin, meaning we are trying to minimize the hinge loss on training sample, resulting in a separator that tries to minimize the number of samples inside the margin, and overfit the separator to the training sample. We indeed get and increasing train error with .
     + The trend of test error should not be increasing or decreasing. The trend should be of that convex function. For lower , as explained above, the separator tries to minimize the number of samples in its margin, overfitting it to the training sample, and not being general enough for the test sample. For larger , the separator minimizes its norm, resulting in larger margin, and the separator is decided by the further samples achieving high error in both train and test samples. The optimal solution should be received with that isn’t too small or too large. We indeed see this trend in the graph, for lower the test error is relatively high, while in larger it’s even higher, and minimum on the test error is received somewhere in the middle.
3. Placeholder.
4. We’ll find a given w that will shatter any combination of labels .  
   Instead of requiring , we’ll demand . Since , then .  
   Now we have d equations . Converting to matrix form we get , where .  
   Since U is a matrix of dimensions with exactly d rows consisting of independent vectors, U must be invertible. So, will separate every possible combination of
5. Placeholder.
   1. We are required to prove .  
      We’ll divide into 3 cases:  
      (1)   
      (2)   
      (3)   
      Therefore, in either of the cases
   2. We need to prove .  
      We know , therefore Proof by induction:  
        
      Base case :  
      .  
      Explanation: Perceptron updates in each iteration with , since , and , then , i.e. only increases/decreases in increments of 1 (or doesn’t change). Since , then it must be at least 1.  
        
      Induction step assuming , proof for :  
      note: last equal is from geometric series sum.  
      We get Thus proving the theorem.
   3. Both previous sections were proven for any (in a) and any (in b), therefore we can take for section a, and for section b.  
      . Combining both:  
      .